Intelligent Archive Visionary Use Case: Virtual Observatories

December, 2003

Robert Harberts, Larry Roelofs
Global Science & Technology, GST
Greenbelt, MD
Harberts@gst.com

H.K. Ramapriyan, G. McConaughy, C. Lynnes, K. McDonald, S. Kempler
NASA/GSFC
Greenbelt, MD 20771

ABSTRACT

Astronomy in the next twenty years will be powered by diverse new observing capabilities. Huge gains in instrument sensitivity covering a broad range of wavelengths for ground and space-based observatories promise abundant data for making unique discoveries, finding solutions to fundamental astrophysical problems, and expanding the frontiers of astrophysical knowledge [1]. These anticipated advances in astronomy will also depend largely on new capabilities for archiving, discovering, accessing, analyzing, and using observational data. Advances in data processing infrastructures will be required to transform accumulated and expected volumes of data into information and knowledge.

For example, the concept of the virtual observatory (VO) anticipates an era of integrated data, information, and services managed in a distributed but cooperative infrastructure. Joining observatory data archives and catalogs into one uniform logical (i.e. virtual) entity with a rich variety of services for discovery, access, and analysis promises to revolutionize astronomy. A virtual astronomy archive, containing observational data gathered in the past, present and into the future, affords a new basis for the science community to "observe" the "sky" through shared digital data and information from desk top computers. New kinds of discoveries and studies become possible because collective observations that include objects cataloged in the entire sky can be indexed across spectrum, time, resolution, and space for instance. In this concept the archive becomes the sky and the digital technology becomes an interface to the observatory.

We explore the nature of future archives by studying the VO concept as a use case scenario. After exploring the various functional capabilities and operations of a VO envisioned by the astronomy community we examine key requirements and implications for archives of the future. We also introduce a new concept of the intelligent archive and examine how its capabilities align with requirements identified in the VO scenario. And with the astronomy use case scenario we identify general characteristics and capabilities that intelligent archives possess for general application in other science enterprise contexts. Finally we discuss key challenges and opportunities for research as an initial "road map" for evolving technology consistent with achieving the goals of the envisioned virtual astronomy archive and observatory.

TABLE OF CONTENTS

Acknowledgement: Special thanks to Dr. Kirk Borne, George Mason University, for his editorial input and enthusiastic contributions to the ideas explored in this paper.

INTRODUCTION

Mapmaking represents a central step-by-step advance in scientific understanding of a subject. Interestingly, astronomers have been methodically mapping the largest "territory" in the history of humankind – the known universe. Piecing together discoveries about the Universe into a "map" or catalog reflects collective knowledge about the universe beyond Earth's atmosphere; knowledge that is rapidly increasing largely as a function of the quality and quantity of observations scientists study. Progressive improvements in scientific technology in the form of supporting research systems (e.g. instrumentation/detectors, computation, applications) and enterprise infrastructures (e.g. observatories, archives, networks) contribute significantly to rapid-pace transformations of our understanding of the universe we inhabit. But celestial "maps" reflect only part of this knowledge.

Survey catalogs constitute another facet of knowledge about the universe that is growing daily. For example, the Sloan Digital Sky Survey (SDSS) is the first large-scale survey developing a detailed digital map and electronic catalog of one quarter of the entire sky. This ambitious project sets the stage for digitally integrating locations of nearly 100 million celestial objects in a large portion of the sky with precise catalog information about the objects [2]. Typically catalogs, such as those for SDSS, contain spatial and object attribute information and photometric emission information for galaxies, stars and quasars. Spectroscopic catalogs contain corresponding information on spectra and identified emission/absorption lines for identified objects [2].

Observational measurements and images, tagged with accompanying catalog information, are retained in archives. As celestial mappings, surveys, catalogs, and scientific papers accrue to the scientific enterprise as whole, the collective results reside in various distributed archives. Over the years astronomy archives have developed and proliferated geographically in number and specialization. But the heterogeneous nature of this collective body of archive data and catalog information reflect progress on one hand and barriers on the other. Few if any standard ways to discover, locate, and access these data by the wider astronomical community exist. Propagation of catalog information updates across different archives can be unevenly synchronized and difficult to maintain. Limitations like these have caused the astronomical community to recognize that significant advantages exist in joining distributed archives and catalogs into one uniform virtual archive.

The concept of a virtual archive stems from the importance of integrating data, information, and knowledge gathered about the sky to address new questions that emerge from discoveries about the known universe. For example, desires to gain further understanding in astrophysics, cosmology, planetary science, astrobiology, and astronomy are driving research emphases in topics pertaining to [3]:

- The nature of matter and energy in the universe
- Events at the dawn of the universe
- How objects like black holes, stars, planets, and large-scale structure form
- How planets form habitats
- The nature of interactions between Earth and its astronomical environment

Future space missions, observatories, and tools to study these topics will add to collective data pools, information, and knowledge in dramatic ways. Common access to these data and multimedia types of information will permit researchers to address the big questions about the

nature of the distribution of matter in the visible universe. Rendering all observations electronically and collectively opens groundbreaking possibilities for making new discoveries leading to profound understanding of the origin and structure of the cosmos.

Concepts and plans are actively being developed to address these expectations. Evidence of this can be found in news and publications describing international virtual observatory projects now established in the astrophysics community (introduced in the Virtual Observatory Scenario section). Data intensive science, like that envisioned for the future of astronomy, prompts new visions for archives that will support 21st century knowledge building enterprises. Intelligent archives represent an example of one such visionary system.

INTELLIGENT ARCHIVE

Future archives will need to manage unprecedented quantities of data efficiently, swiftly, and within constrained budgets [4]. Archives must also adapt and scale capabilities in accord with increasing rates of growth in data acquisition and data product generation. Accomplishing this in a distributed environment of scientists, observatories, archives, and systems is an important challenge facing data management operations. New kinds of automation are essential. Innovative automation techniques need to be embedded throughout an end-to-end process of data transformations. Consequently an intelligent archive (IA) is conceptualized to exist in the context of a scientific enterprise that it supports integrated through a distributed and interoperable cyber infrastructure (see figure 1).

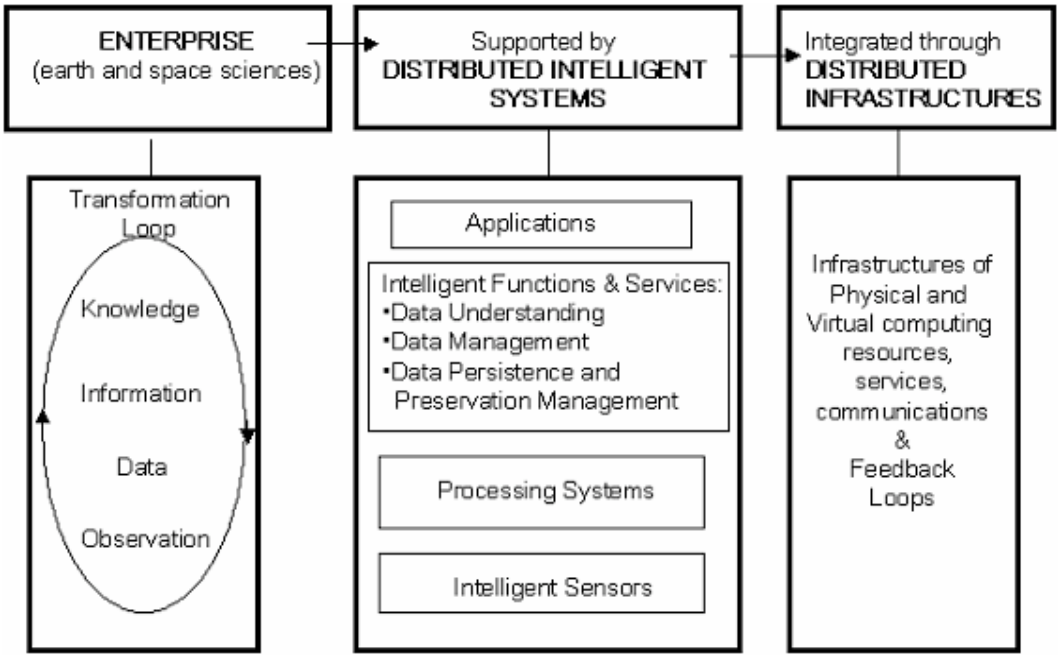


Figure 1: Context of intelligent archives in support of the transformation of observations to knowledge

Developing more intelligent archival functions and services (e.g. data understanding, management, preservation, and distribution) is one technological path for increased automation. Extended archival functionality based on embedded systems intelligence promises to achieve the goal of allowing users to focus more on research and data utilization rather than on data access and processing systems management. The Intelligent Archive study project, supported by NASA Ames CICT, actively explores requirements and concepts for how archives of the future will function and support different science enterprises. We are developing a conceptual architecture for archive systems capable of increasing utilization of data, improving distributed data

management, and automating the transformation of data to information and knowledge [4]. Our research examines ways intelligent systems, such as machine learning, autonomic computing, and algorithms for autonomous and collaborative processes will enable capabilities expected of future operational archives.

Part of the process to develop the architecture for intelligent archive systems involves use case scenarios. Scenarios for use cases help us examine the dynamic nature of envisioned data utilization in applications combined with projections of advances in relevant technologies. In this manner an abstracted architecture can be defined without regard to physical implementation. A generalized IA architecture can then be considered from the perspective of functions needed to support various scenarios. This approach is useful because the functions of an intelligent archive are more stable than the physical architectures and technologies used to implement them. By 'discovering' and abstracting required components and processes from scenarios into functional elements, we are able to explore application strategies of technologies and system resources for future intelligent archives [5].

An outgrowth of this approach is the concept for future capabilities and characteristics of an archive. An IA is differentiated from contemporary archives because the scope of meaning for the term 'archive' is extended from a simple repository of data to one that supports and facilitates derivations of information and knowledge. For example, stored items managed by an IA are extended to include [4]:

- Data, information, and knowledge representations
- Software needed to manage holdings
- Interfaces to algorithms and physical resources to support acquisition of data and their transformation into information and knowledge and storing the protocols to interact with other facilities
- Models to interpret and to maximize the scientific use of the data

The application of intelligent algorithms and intelligent system components enables an IA greater ability to operate more autonomously than conventional archives (e.g. reduced human operator costs). Additionally, an IA is envisioned to provide improved quality of responsive services to users (as an intelligent assistant) with less operator intervention. By embodying new intelligent capabilities an IA can be distinguished from archives of today with regard to its ability for [4]:

- Storing and managing full representations of data, information, and knowledge
- Building intelligence about transformations on data, information, knowledge, and accompanying services involved in a scientific enterprise
- Performing self-analysis to enrich metadata that adds value to the archive's holdings
- Performing change detection to develop trending information
- Interacting as a cooperative node in a "web" of other systems to perform knowledge building (where knowledge building involves the transformations from data to information to knowledge) instead of just data pipelining
- Being aware of other nodes in the knowledge building system (participating in open systems interfaces and protocols for virtualization, and collaborative interoperability)
- Intelligent integration and fusion of data from multiple sources, both local and distributed

As these IA functional capabilities become part of a knowledge building system many benefits for science applications can be imparted. This is inevitable due in large part to dynamic interdependencies that exist between science enterprise applications and IA services. Positive implications for other aspects of the knowledge building system that interact with an IA can similarly be expected. By broadly exploring the interactions various aspects of the virtual observatory concept could have with advanced IA capabilities a clearer understanding of these benefits and possibilities emerge.

In the following sections we examine interrelationships between VO and IA concepts. The first part develops a use case including scenarios that illustrate how some of the envisioned VO capabilities might perform and operate. Operational behaviors expected of a future VO provide insights into requirements, challenges, and goals that can be used to compare with those for an IA. The following case study section provides a comparative assessment of goals and requirements with which to discuss the role an IA has in helping visionary knowledge building systems like the VO for astronomy emerge into reality.

VIRTUAL OBSERVATORY USE CASE

BACKGROUND

Virtual observatories aim to handle a proliferation of large astronomy datasets with built-in software tools for scientists to query and mine data across archives [6]. Over the next decades astronomers expect to practice "precision cosmology" leading to characterizations of the size, structure, and evolution of the universe. Other fields in astrophysics and astronomy similarly expect to pursue research and discovery with digital data and computer intensive tools. Achieving goals such as these will require the collection and integration of petabytes of data from space and ground surveys [7]. Already several projects worldwide are developing virtual observatory mechanisms to federate collections of data and information for an entire scientific discipline (e.g., National Virtual Observatory, Astrophysical Virtual Observatory, Astrogrid, Astrovirtel). The goal is to knit these projects together so that ground and space-based astronomy archives are linked and accessible to all (see figure 2). Intelligent archives are well suited to the functional task of integrating services that a virtual observatory requires to manage and make accessible astronomy datasets. We will discuss this point further in succeeding sections.

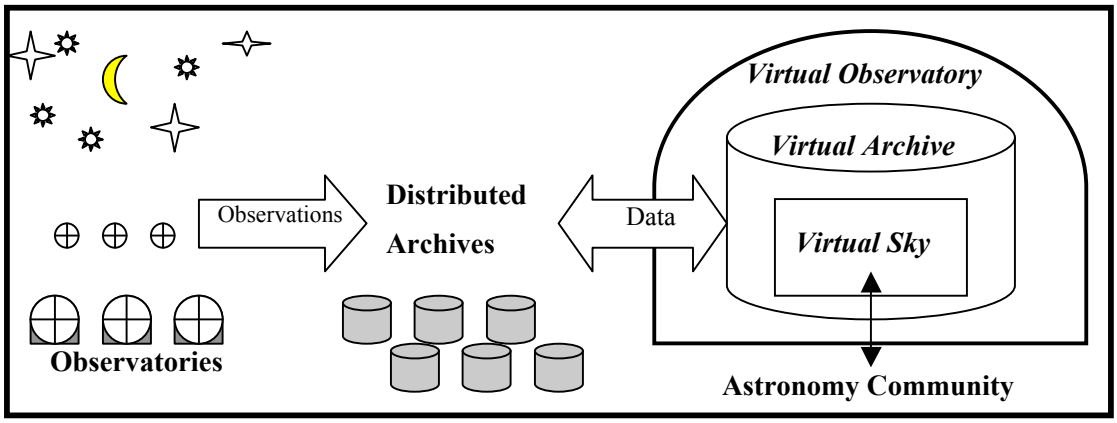


Figure 2 – Context for collecting observations into a virtual sky utilized by astronomers

DESCRIPTION

The data avalanche in astronomy is attributable to increasing use of Charge Coupled Device (CCD) cameras for telescopes. Advances in electronics permit the deployment of gigapixel instruments used to conduct multi-spectral surveys of the sky. The volume of pixels from a typical astronomical CCD detector doubles every two years with this rate increasing in the near future [8]. Assuming that low-light astronomical CCD arrays deployed in telescope focal planes follow Moore's Law (quantity doubles roughly every 1.5-2 years) then pixel counts will continue to grow from 10e8 pixels in 2000 to 10e10 pixels by 2010 and 10e11 by 2020 [9]. Gigapixel instruments that can generate 100 terabyte surveys (such as the Sloan Digital Sky Survey) will easily overwhelm current archiving capabilities. Like genomics and other cutting-edge fields of science, astrophysics must cope with the growing problem of how to manage and make use of enormous distributed amounts of data generated by digital-based instruments and experiments.

A virtual observatory therefore, is a science-driven effort that emphasizes bottom-up sharing through common integrated services for both observational data and toolkits of analysis software. Many existing astronomical data archives are associated with a specific instrument. Each specific archive has historical reasons to archive data in a particular way. But systematic studies and surveys of the celestial objects depend on combining multi-spectral and multiple instrument data from multiple archives. VO software, standards, and supporting infrastructure (i.e. grid) are intended to integrate and interoperate physical archives into a virtual archive (see figure 2). The astronomy community then interacts with the data that has been collected and unified into a "virtual sky".

Archival institutions participating in the National Virtual Observatory advocate one solution with regard to the virtual archive/VO concept. In this solution, each institution maintains control over individual data holdings but institutions can share data by conforming to extensible metadata standards and interchange protocols [10]. Users in the community will also be able to share data and analysis tools through properly designed virtual observatory interfaces. A core set of standards, interoperability, and management services will be necessary to support distributed virtual observatory operations.

OPERATIONS

Different telescopes acquire different observational data in different formats managed in geographically distributed archives. Many archives also link data from multiple experiments. A virtual observatory of collective archives will operate based on a common system approach for data pipelining, archiving, discovery and retrieval. It also ensures easy access to data in these archives by a diverse community of users. The system will enable distributed development of a suite of commonly usable new software tools for querying, correlation, visualization, and statistical comparisons of shared data.

Virtual archives utilize high-speed network connectivity among participating archives and terascale computing facilities [11]. Because bandwidth is a limiting factor computation needs to be performed close to the data. But Grid technology infrastructures will allow remote access to both data and computing/analysis facilities facilitated by an operational virtual observatory.

SCENARIOS

A clearer understanding of the nature of virtual observatories can be gained from scenarios that describe facets of their operational characteristics or behavior. They also help "illustrate", as glimpses, the capabilities a virtual observatory, virtual archive, or virtual sky would offer the science community in early 21st century decades. Along these lines three sample scenarios are described below to support further exploration of VO and IA goals and requirements in following sections.

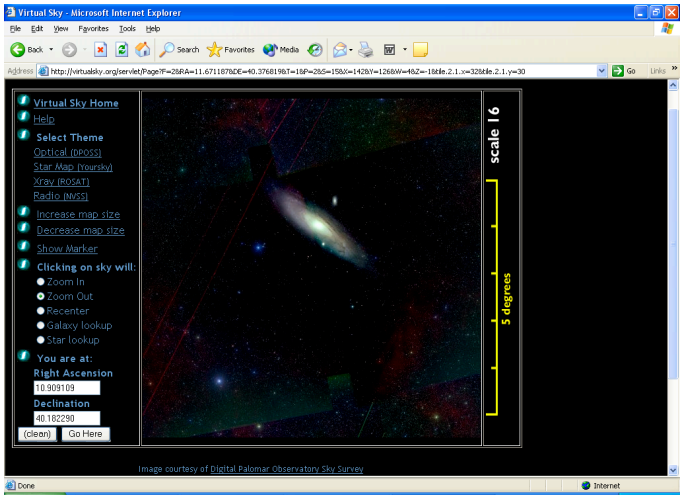
REAL-TIME OBSERVATION AND ARCHIVE COLLABORATION – SCENARIO 1

In a distant region of space a 900 millisecond flux of gamma rays announce the death of a massive Wolf-Rayet star. After traveling many light years the flux of gamma-rays are detected by Earth-orbiting gamma-burst scanners. The coordinates for the source of the gamma-rays are immediately distributed to ground-based and space-based observatories with a priority flag. Intelligent systems that are automatically monitoring an array of observation requests, alert bulletins, active, planned and scheduled observation plans, broker new observing priorities with available online observatories. Three robotic observatories swing into action using instructions and coordinates brokered by the virtual observatory and make new observations in several wavelengths. The virtual observatory also negotiates with several other observatories to revise previously scheduled observations in time to orient radio telescopes with coordinates for the gamma-ray burster. Data streams that flow from each of the observing telescopes into the virtual observatory archive trigger a set of priority notifications for issue via a subscription service. The subscription service matches the priority message with a pre-registered list of scientists to be notified about the event and available observation data. While the science community is alerted the death of a star and the birth of a massive new black-hole have been recorded and cataloged nearly real-time.

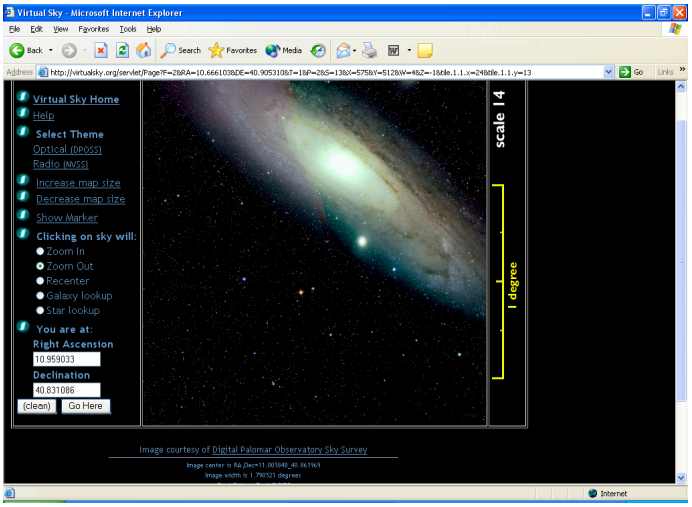
In this scenario highly interoperable and integrated systems interconnect observational instruments, various data management services, and the user community. A high degree of automation enables responsiveness in the system required to detect and rapidly transform observations into useful information and data. What is unclear but implied in this high level scenario concerns the chain of services and capabilities required to achieve near real-time responsiveness. The intelligent archive study has explored this territory by examining two-way communications through end-to-end processes connecting “sensor webs” with archives and user software. IA relevance to functionally integrating “sensor webs” of observatories into a virtual observatory will be discussed further in the VO and IA Crosswalk Study section.

ARCHIVE IS THE SKY – SCENARIO 2

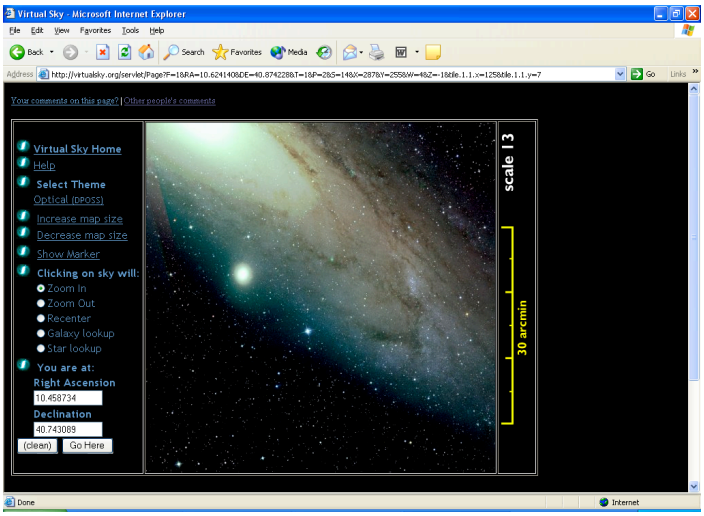
The virtual archive interacts directly with observatories, archives, computing infrastructure resources, and thousands of professional and amateur astronomers worldwide. High speed networked connectivity permits interactions among these elements. A regular stream of new observations is added to those accumulated for years, even decades, in the form of registered data in the virtual archive. These data are automatically classified and “mapped” into a multidimensional mosaic of the whole sky such that all observations can be found by referencing a portion of the sky and zooming in with a variety of interface tools (see examples 1-3) [12]. These tools perform like a virtual telescope but with the advantage of instant access. In fact the virtual archive presents its holdings to astronomers as a virtual sky that can be “observed” unconstrained by weather, scheduling, remoteness, time, budget, time of day, or time zone.



Example 1: 5 degree zoomed observation of a galaxy



Example 2: 1 degree resolved observation



Example 3: 10 arc minute resolution

The virtual sky represents a unified and standard portal to collective but multiple resolution observations on the sky on the one hand and the unification of data archives and databases on the other. Professional and amateur astronomers "see" the sky through observational data collected by multiple observatories organized into one virtual archive. The collective archive is the sky in the sense that research observations can now be made by querying observational data from a unified archive, not just from a telescope.

Astronomers worldwide connect to the virtual archive to issue requests for observations and to collect observations of interest. An all-sky atlas, for instance, guides observers through gigapixel arrays with arc-second resolution of multiple wavelength observations to find objects. This kind of atlas constructed from and applied to unified datasets gains success because astronomers often look at the same objects on the sky and because the time scale of changes is relatively slow for most objects [13]. An interactive zoom-able 3-D interface into virtual archive data resources allows astronomers to conduct high fidelity, cross-archive, interactive exploration of the sky from their personal computers.

PERSONAL COMPUTER IS AN OBSERVATORY – EXAMPLE 3

The virtual observatory (i.e. VO@home) puts at the finger tips of a scientist the ability to enter a general or specialized query based on a set of parameters to discover objects or outliers from data collected over the entire sky [14]. For example, using a laptop a cosmologist quickly connects to a VO portal customized according to personal preferences. These custom views are selected for data subscriptions, data processing/mining services, analysis tools, galaxy atlases, observation scheduler, alerts, collaborators, digital library, and data viewing menus. The VO portal to the virtual archive provides uniform, high-quality photometric and spectroscopic observations of millions of objects available in digital form and format choices that are customized for desktop or laptop use by every astronomer [2]. VO@home, for instance, can also use the massive distributed computational power of home/office PCs to mine the petabyte data stream for such things as comets, asteroids, supernovae, variable stars, and rare events.

In a parallel scenario astronomers at home can register and plug in their robotic, Internet connected telescopes into the VO. Thus connected the robot telescope performs as a smaller version of an autonomous, semi-intelligent observatory [15]. While connected it can be accessed and tasked by the VO to ingest observations. These observations can be assimilated into a data-intensive study of near Earth objects for example. In turn, the @home astronomer (i.e. amateur) can simultaneously access the virtual archive using interactive virtual sky services to receive data to augment the amateur's personal project. Two-way interactivity and multi-mode VO operation permit observatory work to be conducted through personal computers. And, in this case, @home astronomers can participate in collective observation campaigns by connecting a smart home robot telescope to the VO.

GOALS AND CHALLENGES OVERVIEW

The three previously described scenarios encompass an overall end-to-end goal where observatories and archives are collectively linked so powerful on-line research becomes possible to a wide user community. A representative summary of scientific aims underscoring this goal for VO projects have been formulated to: [16]:

- Improve quality, efficiency, ease, speed, and cost-effectiveness of on-line astronomical research
- Make seamless and transparent integration of data
- Remove data analysis barrier to interdisciplinary research

- Make manipulation of large datasets easy and powerful

As the scenarios evince, a primary goal actively advanced by many astronomers is uniting the telescopes and observatories of the world through the data they produce. Since observational data flows through many processes for acquisition, ingest, validation, and data management into various distributed archives, an underlying challenge involves uniting the archives. When this happens a virtual observatory will emerge. And when the VO becomes a reality it will be the tool for a new era of astronomy consisting of many terabytes of survey data collected largely by CCD-enabled telescopes around the world (including space-based observatories). Hence another important challenge associated with this goal for the astronomy community concerns how to cost-effectively manage and efficiently use the tsunami of data threatening to overwhelm astronomers. Echoes of this challenge ring within another high-level goal.

VO initiatives are principally oriented to improving access to abundant data by the science community. For instance the chief architect of the National Virtual Observatory has stated that the goal of NVO is to make sure that "the current generation of professional and amateur astronomers is not overwhelmed by the chores of getting the actual data". [17] A joint NASA/NSF Science Definition Team similarly stressed that the virtual observatory must make practical those studies that today are too costly in terms of the resources required to acquire and use data [18].

Realizing significant reductions in overhead associated with data access for humans using advances in information technology remains a perpetual challenge in the foreseeable future. Achieving "easy access" to data resources appears to be strongly related to the challenge of how system complexity and complicated processes can be hidden from the user. These kinds of challenges are applicable to efforts for advancing system automation, optimization, interoperation, and functional capacity scaling [19]. Various categories of requirements emerge because these challenges span different layers of the system architecture, applications, and applicable standards.

SAMPLE SET OF GLEANED VO REQUIREMENTS

Given an overview of high level goals and associated challenges we provide a sample of requirements identified for VO capabilities. VO requirements such as those below can be related to key categories of IA requirements for comparative study.

- Seamless mosaic of the sky – whole-sky mosaic of observational data
- Catalog overlays on observations
- Automated discovery and usage techniques
- Data subscription and event alert notification services
- Quality standards and validation process rapidly and accurately optimized
- Bulk processing of data via distributed analysis engines
- Data sub-setting, reduction, and distillation capabilities
- Software needs to directly communicate with the data warehouse
- Move the code, not the data for processing
- Share data within a common framework of metadata standards
- Join diverse sets of catalogs
- Indexing of very large databases

- Middleware that couples archives to users
- Couple models to archives
- Scalable computational and storage – parallelism in processing and storage
- Couple computational resources to data and analysis resources (e.g. grid)
- Operational structure of distributed databases
- Resource discovery tools – astronomical query language (AQL) and web services
- Ambitious kinds of on-line analysis – complex query handling – analyze in situ
- Query estimators to measure the same size of the returned results in advance
- Collect multi-wavelength data on objects cross-matching optical, IR, radio catalogs
- Data intensive science – search for rare objects; enable discovery in multi-parameter space
- Calibrate, classify, index, register, and archive data in formats so many scientists can use them easily
- Caching and replication of "popular" data and data products

VO AND IA CROSSWALK STUDY

KEY CATEGORIES OF IA REQUIREMENTS

Having explored science-driven but technology enabled requirements, goals, and challenges for astronomy in the next twenty years we turn our attention to a crosswalk between those for VO and those for IA. The objective is to integrate distinct but complementary requirements into a generalized viewpoint for future archives. Requirements reflect the nature of the conceptualized system as well as the goals of the science community it will serve. A crosswalk is a technique for navigating between various differing perspectives on goals and requirements by analyzing and mapping the similarities of each [20]. We will use key categories of requirements that the IA study team identified to anchor a comparative assessment.

Presently there are five high level categories of IA requirements that encapsulate capabilities that an IA should exhibit around 2015. A brief characterization of each and a discussion about relationships with VO requirements, insights, and perspectives follow.

KNOWLEDGE-BUILDING SYSTEMS

An IA-supported knowledge building system is conceptualized to consist of many (~200) archives collectively managing petabyte magnitude data holdings. Collective archives that support a particular scientific enterprise exist in a robust collaborative environment of 'nodes' providing on-demand chains of services. Knowledge building systems will be capable of performing intelligent data understanding (IDU) and data mining on cross-discipline multi-decade datasets from ~6 archives within time periods relevant to specific applications. An example might include mining long-term holdings for trend analysis. Others could include processing IDU algorithms 24/7 on accumulated data as well as on ingested data streams. Results of near real-time monitoring of new data and progressive mining of cumulative data archives enrich catalogs, metadata, and other information resources.

APPLICATIONS OF SCIENCE DIRECTLY USABLE

This category emphasizes requirements for an IA to support scientific applications with highly efficient operational capabilities. These capabilities include a self managing robust system exhibiting "lights out" or self-healing (e.g. autonomic) operations, reduced latency in computation, intelligent resource optimization, and autonomous but rapid quality assessment (QA) in-line with data product generation. Quality science data can be processed rapidly (minutes) in a highly automated, accurate, self-learning (adaptive) system environment where resources are dynamically matched to production parameters.

INTELLIGENT ALGORITHMS EMBEDDED IN CYBER-INFRASTRUCTURE

The intelligent cyber-infrastructure is a required science-enabling communications and software infrastructure. A primary requirement of this infrastructure is that it contains scalable, robust, and intelligent algorithms capable of supporting a particular scientific enterprise and its operational processes. This is characterized by distributed on-demand data generation using adaptive algorithms, concurrent processing coordinated in multiple service-chains and production timelines in minutes for selected service chains. Embedded intelligent algorithms enable distributed services to be identified, configured, an engaged in meeting operational service demands generated by thousands of members of the science enterprise. One example is a distributed service chain for data mining of entire datasets for features, phenomena, events, objects, and pre-cursor statistical signatures.

STRONG INTEROPERABILITY FABRIC

Requirements for strong interoperability among systems align with goals for achieving unified interchange among distributed enterprise system resources, services, and access mechanisms. A strongly interoperable fabric is a pervasive mesh of common standards, tools, software, and protocols for systems and applications to use as an interface. The fabric supports automated data registration, discovery, and utilization among collaborative systems for example. Requirements for an interoperability fabric also address the need for archives to provide improved semantic access to diverse data and metadata (i.e. incorporating a fully developed ontology in each archive). Knowledge representation, ontologies, and semantic web are implicit in achieving greater semantic access to data and support services. Meeting requirements like these establish advanced interoperability capabilities at various levels (i.e. functional, operational, and semantic) of the fabric.

FEEDBACK LOOP

This category of requirements focuses on incorporating links between an IA and data acquisitions resources on one hand and knowledge building processes (i.e. modeling) on the other. Feedback consists of increasing "awareness" within the overall system about such things as states, patterns, conditions, opportunities, events, etc. with which to progressively enhance performance and interoperable responsiveness. For self-analysis and autonomous anomaly response capabilities a feedback loop incorporates unsupervised algorithms for multiple anomaly detection for rapid operational awareness and response. Intelligent data understanding algorithms would be embedded in all mission data production processes including capabilities for scanning 24/7 all long term-archives for discoveries to inform scientists via alert lists. Additional feedback loop capabilities include: automatic filtering and risk assessment of commands issued to space and ground-based observation platforms, brokering resource availability and request queues 24/7, and rapid adaptive scheduling between requests from modeling systems and responses to the model from IA services.

Note: While we have identified high level requirements for an IA we recognize that our vision of intelligent systems, configured into knowledge building systems for different science enterprises,

also contain as yet undiscovered requirements. So too is the prospect that eventual development and implementation of these systems possess unimagined benefits. Proof-of-concept will ultimately be realized in practice as these systems transition from research prototypes to systems of service. At this point a cursory evaluation of entangled IA and VO concepts offers a preview of a knowledge building system.

SYNOPSIS OF RELEVANCE AND BENEFITS

The five high level IA requirement categories and VO goals (see Table 1) can be compared for insights into slotting IA–related components into an overall VO knowledge building system (see Figure 3). Characteristics of a knowledge system are consonant with the goals and requirements for VO projects. For example, a VO must address the need to handle the explosion in size of astronomical datasets delivered by powerful new instrumentation and ambitious whole-sky surveys for the whole astronomical community. But currently distributed astronomy archives have great scientific value going unexplored and underexploited in large unconnected datasets in astronomy [21]. VO and IA concepts correct this by unifying data resources into a shared or virtual archive. The virtual archive is supported with interoperating services that apply collective computational resources to the task of transforming data into information and knowledge. Astronomy will benefit from an operational knowledge system capable of intelligent data understanding, collective long-term dataset mining, and cross-dataset/cross archive analysis for increased scientific return.

IA Requirement Categories	**VO Goals**
• Knowledge building system • Applications of science directly usable • Intelligent algorithms embedded in cyber-infrastructure • Strong interoperable fabric • Feedback loop	• Improve quality, efficiency, ease, speed, and cost-effectiveness of on-line astronomical research • Make seamless and transparent integration of data • Remove data analysis barrier to interdisciplinary research • Make manipulation of large datasets easy and powerful

Table 1: High level comparative requirements and goals

The virtual archive of shared, seamless, and transparent integration of archives and databases must provide faster quality services than experienced with individual archives of today. Therefore IA requirements to support scientific applications with very efficient operational capabilities at the archive level and for strong interoperable fabric for unifying distributed system resources at a collective level align directly with each VO goal. Manipulating large datasets and conducting on-line analysis and astronomical research through the VO involves a host of services. These services range from science application/science user to machine level or infrastructure. Automation and intelligent systems are critical to the success of quality, speed, efficiency, and cost-effectiveness. Embedded intelligent algorithms are highly applicable to IA and VO strategies for success. Feedback for example allows for new observations to be requested based on the archival data analysis. Continued research, however, will be required throughout all phases of development and sustaining engineering to realize VO/IA goals. The feedback loop will help inform opportunities for research to address deficiencies and technology infusion for improvements. A feedback loop is an essential adaptive component for evolving VO system capabilities with the knowledge building objectives of the astronomy community.

IA and VO concepts for unifying and integrating various science enterprise resources, (e.g. sensors, data, computational services, data exploitation applications, etc.) are complementary. Physical astronomical observatories (e.g. glass, CCD, radio) are comparable to sensor webs and observational platforms described within the context of the IA architecture [4, 5, 22, 23]. Treating sensors or observatories collectively as a resource to integrate with other integrated system resources constitutes the basis for "virtualized" physical resources [4, 24]. Benefits of virtual observatories, virtual archives, and even virtual skies for the science community are traceable to:

- Plugging observatories into an interoperable layer for two-way broker of observation data acquisition and tasking requests
- Creating systems that are highly responsive to on-demand requests for quality data
- Increased scientific productivity and cost effectiveness of resource sharing
- Tools for simultaneous browsing of multiple archives and visualizing results
- "Democratization" of resource discovery, access, and utilization
- Gathering a large body of data accurate enough to address for the first time a broad-range of fundamental questions
- Knowledge gained from interdisciplinary research based on shared collective data

For heterogeneous and distributed resources to benefit a community of astronomers applicable VO and IA systems can be organized into an interoperable knowledge building system. Figure 3 depicts the context model for a VO knowledge building system and its underlying elements, drawn from the IA study. In this model an interoperable "fabric" of standards, interfaces, and services (including network and computational infrastructure services) implement the VO as a logical interface for astronomers.

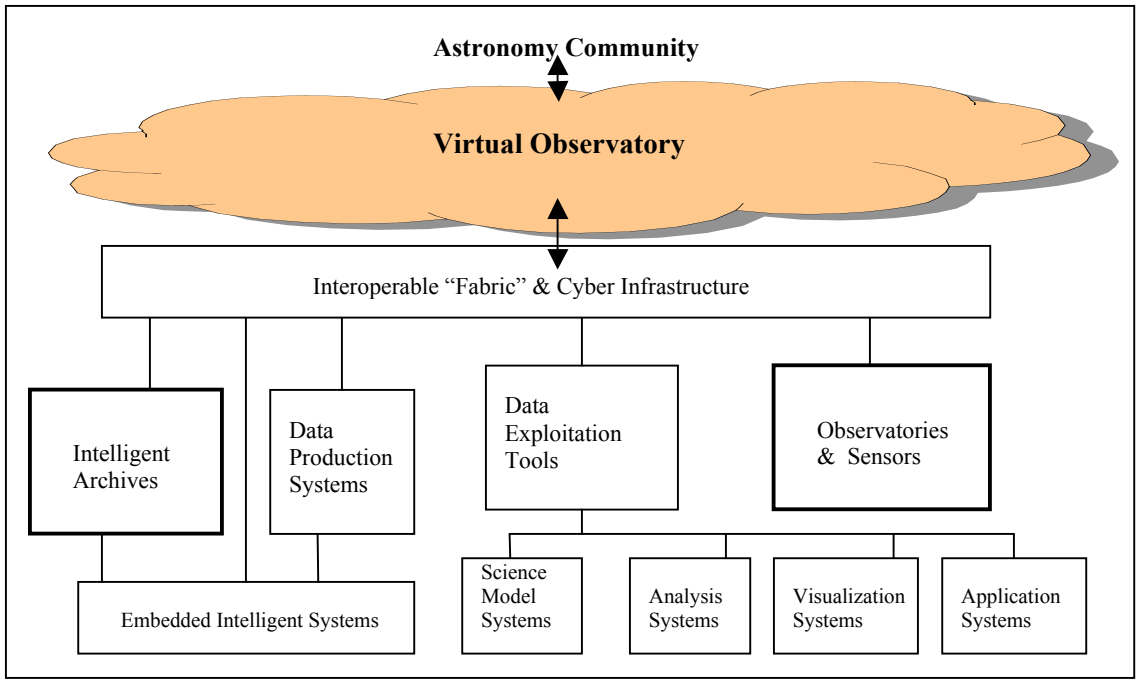


Figure 3: IA context model of underlying elements for a virtual observatory

Core elements for supporting the VO also plug into the interoperable infrastructure layer. These elements include the intelligent archives, data production systems, data exploitation tools, and the observatories or sensors used to make digital observational measurements. Integration of these elements (and other underlying systems including modeling, analysis, visualization, value-added applications) through registration/interfacing on the interoperable layer enable distributed systems to participate in dynamically composed service-chains. Services can be "chained" in different combinations depending on a service request. Service requests can be brokered for any interoperable system element as well as for top-level user applications.

System services either chained or standalone in this context can also be encapsulated as high-level capabilities for the user community thus hiding complexity. Advertised ease-of-use capabilities like custom data mining, analysis, or mosaics of n-dimensional data products, invoke automated but flexible, configurable distributed services for fulfillment. Similar value for astronomers is gained from query and look-up capabilities that leverage system intelligence with chained services to return data content information. Capabilities like these spare the user substantial overhead.

OBSERVATIONS

The VO case study has shown that this science enterprise is generally composed of:

- Large data collecting systems (i.e. observatories) - those with sensitivities and resolutions to perform detailed observations of objects and phenomena
- Large data management systems (i.e. archive/repositories) – those that acquire and collect data in forms that can be stored, managed, made discoverable, and used by scientific processes for deriving information and understanding
- Large knowledge building systems – people, organizations, scientific questions and investigation drivers, missions, techniques, tools of analysis and computation, algorithms, processes, methods, literature, and discourse for deriving information and transforming it into understanding

These main features of the VO are representative of other science-based enterprises considered by the IA study team (e.g. earth science, weather forecasting, precision agriculture) [4, 5, 24]. Each discipline shares similarities in challenges for the future with regard to (1.) managing increasingly large data resources generated by advanced sensor and observing technologies and (2.), maximizing the use of these data via timeliness and in-line processes for intelligent data understanding and automated management.

The intelligent archive concept addresses the challenge of full utilization of data. Furthermore, an IA represents an integral component in the transition from data and information systems to knowledge building systems. In a knowledge building system architecture observational systems, data management, archiving, science processing, and automated intelligent data understanding interoperate in a tightly integrated networked infrastructure. Collective capabilities of this envisioned knowledge building system are proposed to assure that massive volumes of data, both new and accumulated, are utilized fully [22].

CONTINUED DEVELOPMENT

ROLES FOR SYSTEM INTELLIGENCE

Given the goals for envisioned virtual observatories to better serve the knowledge building science community, VO implementation designs present interesting challenges for computer science and information systems technology to address. Many challenges concern how to establish the kind of capabilities necessary to unify data, information, and services without adding more overhead to human work loads and costs. More so, a paramount challenge concerns learning how to maximize scientific data utilization in a distributed environment where observational data resources are increasing dramatically. These challenges have a bearing on the kind of intelligent systems applications that are developed with demonstrated abilities to:

- Significantly increase scientific return
- Substantially reduce effort scientists expend on discovering and preparing data for research
- Lowering the cost of data management over time
- Increase quality and accuracy of data
- Decrease latency while increasing responsiveness to requests

Opportunities for intelligent systems in virtual observatories include: automated data corrections, reprocessing, and recalibrations (i.e., data "curation"); cross-dataset and cross spectrum data mining with automated phenomena "discovery"; open-ended resource discovery; enriched cross-archive metadata generation for uniform query and browse, data fusion and combination management. Roles for intelligent archive capabilities are aligned with the goals for virtual observatories to closely link and coordinate data curation, archive data management, uniform access, and data mining services. Overall intelligent archive and Grid functions will allow astronomers to more easily find, access, analyze, and manipulate data as if the data were local to the astronomers' workstations.

TECHNOLOGY CHALLENGES AND OPPORTUNITIES

Concepts for multi-wavelength synoptic virtual observatories have already inspired active research in Grid technology. Grid infrastructures will enable large-scale interoperable sharing of data and computing resources necessary to support next generation astrophysical surveys, analysis, and research. In addition to grid technology various other challenges for developing new VO infrastructure layers include or embed innovative algorithms and tools needed for data mining, statistical analysis, visualization, data fusion, data filtering, and virtual interactive interfaces. Infrastructures capable of supporting space science communities will require new protocols and standards for transparent access to multiple distributed heterogeneous systems. These distributed systems will need to be developed with capabilities that support such things as data discovery, metadata and query interchange, code-shipping, and data product delivery. Achieving sophisticated capabilities such as these will also depend on developing intelligent interfaces required for storing and managing both data and metadata throughout a virtual observatory.

Opportunities exist for applying intelligent archives' concepts to this emerging but visionary space science system environment. Some of the first steps needed for such an application are: identifying technical requirements, mapping the requirements into the various components of the system, and for each component, defining the technical interaction scenarios with other components of the system. These initial steps will reveal with increasing specificity requirements and hence the opportunities to meet them with innovative technologies and technical solutions mapped to where they have the most value in such a system.

SOME VO CHALLENGES FOR IA

VO conceptualizations contain some deeper challenges for IA and knowledge building systems to study. Performance, process, and service requirements for creating a "virtual sky" out of configured interoperable elements merits technical assessment. Communication pathways and processes among collaborating elements (e.g. observatory, archive, and catalog) pose interesting challenges for system level process management especially in real-time priority situations. Variants of this challenge, applied to other kinds of service-chaining, pertain to:

- use cases involving acquisition of new observational data, relating the new with old data, making use of recent data, mining old data, and fusion of different data types from different times over same sky coordinates.
- embedding intelligence in observatory robotics, operations, and interfaces
- producing mosaics of disparate data on-demand vs standard data production
- easy query look-ups for data and data content (objects, phenomena, interesting or anomalous features, etc.
- self-learning from query/result streams with regard to optimizing searches, when to cache query results, how to associate valuable links to related information or data, relationships among objects in the corpus of data, how to make recommendations to users for new discoveries, and close knowledge feedback loops between publications and references to data resources used

VO concepts also serve as a basis for good test cases to refine development of the IA and knowledge building system architecture. The VO case study affords operational scenarios with which to design a near-term IA architecture and application. This testbed study would be useful for discovering bottlenecks, interface complexities, optimization pathways, opportunities for intelligent systems integration, and help confirm/validate specific design concepts.

CONCLUSIONS

In this use case study virtual observatory, virtual archive, and virtual sky have been introduced as concepts for enabling and supporting next era astronomical research. We also introduced the concept of the intelligent archive and its role within the context of a knowledge building system. Exploration of how these complementary concepts align as requirements for new capabilities for systems to support science provide insights into tangible configurations, architectures, and technical research pathways for development.

The transformation of observational data into knowledge occurs in a distributed environment of people, systems, and infrastructure. We have considered the role of future archives with regard to VO goals and requirements in developing concepts for intelligent archives. In this paper we also discussed core IA concepts that relate to data and information management issues for astronomy.

As all-sky surveys become technically feasible, 3D maps of the universe will emerge within the next 15 to 20 years. Deployed 10e11 pixel CCD enriched telescopes arrayed in ambitious detector architectures for example, promise huge data resources that advanced analysis techniques will transform into resources like ultra high resolution 3D cosmic atlases [9]. Developments such as these will enable much of astronomical and cosmological research to increase our understanding of such things as our own origins, and the evolution of matter and energy to the present, the constituents of the Universe, and the processes that shape the cosmos and its components.

REFERENCES

[1] Najita, J., Strom, S., "Science Enabled by a 30-m Telescope", Future Research Direction and Visions for Astronomy, Dressler, A., ed., Proceedings of SPIE, vol. 4835, August, 2002, p. 2.

[2] Szalay, A., Kunszt, P., Thakar, A., Gray, J., Slutz, D., Brunner, R., "Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey", Microsoft Resarch Advanced Technology Division, Technical Report MS-TR-99-30, Feb., 2000, pp. 1-12.

[3] Kilston, S., Bally, J., "Potential paths in Space Astronomy Over the Next 50 Years", Future Research Direction and Visions for Astronomy, Dressler, A., ed., Proceedings of SPIE. August, 2002, Vol. 4835, pp. 98-109.

[4] Ramapriyan, H.K., McConaughy, g., Lynnes, C., Kempler, S., McDonald, K., Harberts, R., Roelofs, L., Baker, P., "Conceptual Study of Intelligent Archives of the Future", Report prepared for the Intelligent Data Understanding program, http://daac.gsfc.nasa.gov/IDA/IA_report_8-27-02_baseline.pdf.

[5] Harberts, R., Ramapriyan, H.K., McConaughy, g., Lynnes, C., Kempler, S., McDonald, K.., Roelofs, L., "Intelligent Archive Visionary Use Case: Advanced Weather Forecast Scenario", white paper for the Intelligent Data Understanding program, July, 2003, http://daac.gsfc.nasa.gov/IDA/ posted draft pending.

[6] Fender, T., "Astronomers Envision Linking World Data Archives", Physics Today, February, 2002, p.20.

[7] National Research Council, "Astronomy and Astrophysics in the New Millennium", National Academy Press, 2001, Washington, D.C.

[8] Szalay, A., Gray, J., "The World-Wide Telescope", Science, Vol. 293, September 14, 2001, p. 2037.

[9] Bally, J., Kilston, S., "Mapping the Entire Universe: A Goal for Gigapixel Arrays, 3D Spectro-Imagers, and Large Telescopes", Future Research Direction and Visions for Astronomy, Dressler, A., ed., Proceedings of SPIE, August, 2002, Vol. 4835, p.34.

[10] NVO Interim Steering Committee, "Toward a national Virtual Observatory: Science Goals, Technical Challenges, and Implementation Plan", Virtual Observatories of the Future, ASP Conference Series, Vol. 225, 2001, pp. 353-372, http://arxiv.org/abs/astro-ph/0108115

[11] Brunner, R.J., Djorgovshi, S.G., Szalay, A.S., (eds.), Virtual Observatories of the Future", Astronomical Society of the Pacific Conference Series, Vol. 225, 2001, San Francisco, CA, p. 357.

[12] Web portal to images of the night sky, http://virtualsky.org/index.html

[13] White paper: "A national Virtual Observatory for Data Exploration and Discovery", California Institute of Technology, October 21, 1999, p.3; http://channelonline.itpapers.com/abstract.aspx?scid=392&kw=&dtid=0

[14] Borne, K., NASA GSFC/GMU, personal communication, 2003.

[15] Britt, R.R., "Automatic Astronomy: New Robotic Telescopes See and Think", Sapce News, June 4, 2003.

[16] Lawrence, A., “AstroGrid: Powering the Virtual Observatory”, Virtual Observatories, Szalay, A., ed., Proceeding of SPIE, vol 4846, August, 2002, pp. 6-12.

[17] Schechter, B., “Telescopes of the World, Unite! A Cosmic Database Emerges”, The New York Times, May 20, 2003, http://www.nytimes.com/2003/05/20/science/space/

[18] Hanisch, R., “Building the Framework of the National Virtual Observatory: Status Report”, Virtual Observatories, Szalay, A., ed., Proceeding of SPIE, vol 4846, August, 2002, pp. 13-19.

[19] Morse, H.S., Isaac, D., Lynnes, C., “Optimizing Performance in Intelligent Archives”, NASA, GSFC, January, 2003, http://daac.gsfc.nasa.gov/IDA/

[20] Baca, M., ed., 1998, Introduction to Metadata Pathways to Digital Information, Getty Information Institute, p. 19.

[21] Quinn, P, Benvenuti, P., Diamond, P., Genova, F., Lawrence, A., Mellier, Y., “The Astrophysical Virtual Observatory (AVO): Progress Report”, Virtual Observatories, Szalay, A., ed., Proceeding of SPIE, vol 4846, August, 2002, pp. 1-5.

[22] McConaughy, G.R., McDonald, K. R., “Moving from Data and Information Systems to Knowledge Building Systems: Issues of Scale and Other Research Challenges”, NASA, GSFC, July, 2003, http://daac.gsfc.nasa.gov/IDA/

[23] Harberts, R., Ramapriyan, H.K., McConaughy, G., Lynnes, C., Kempler, S., McDonald, K., Roelofs, L., “Intelligent Archive Visionary Use Case: Precision Agriculture Scenario”, NASA, GSFC, August, 2003, http://daac.gsfc.nasa.gov/IDA/

[24] Harberts, R., Ramapriyan, H.K., “Concepts for Models and Virtualization in Future NASA Science Enterprise Systems, IASTED International Conference: Applied Modeling and Simulation, AMS, Cambridge, MA, November 4-6, 2002.